Faculty Working Papers

THE USE OF CONFIGURATIONS OF PURCHASE LIKELIHOODS TO PREDICT AUTO PURCHASES

Ellen Liebman and Robert Ferber

#385

College of Commerce and Business Administration
University of Illinois at Urbana-Champaign

College of Commerce and Business Administration

University of Illinois at Urbana-Champaign

March 10, 1977

THE USE OF CONFIGURATIONS OF PURCHASE LIKELIHOODS TO PREDICT AUTO PURCHASES

Ellen Liebman and Robert Ferber

#385

The Use of Configurations of Purchase Likelihoods to Predict Auto Purchases

ABSTRACT

Using consumer panel data, this paper explores the extent to which automobile purchases can be predicted by functions containing purchase likelihoods for automobiles and for other durable goods both in the very recent past and going back further. Despite low goodness-of-fit, predictions of auto purchases are found to be quite high when a logit model is employed.

The Use of Configurations of Purchase Likelihoods to Predict Auto Purchases

Ellen Liebman and Robert Ferber

Introduction

The work reported here is an outgrowth of a previous empirical exploration using data from the first two waves of the Illinois-Berkeley panel of young married couples in Peoria and Decatur, Illinois. These couples had been married in the summer of 1968, and the husband was 30 years of age or less at the time. The couples (313 at the start) were interviewed that fall and approximately every six months thereafter.

The earlier analysis had both substantive and methodological concerns. Our substantive concern was with the investigation of whether the configuration of reported likelihood to buy 13 different durable goods at one time showed any relationship to the purchase of one of those durables, automobiles, six months later. Likelihoods were obtained as subjective probabilities on a scale from 0 to 100. The focus was on auto purchases in view of their magnitude and of their high frequency of purchase.

Implicit was the hope that the likelihood variables would absorb enough of the complex of factors involved in deciding to make a purchase, that they and the constraints arising from their interactions would yield enough information to predict purchases. This assumption clearly must vary in its validity with the amount of time covered by the likelihood assessment. Thus, assessments made today will function better as indicators of the factors relevant to what will be done tomorrow than they will for what will be done next month, etc. Since the likelihood questions covered a horizon of six months, and sometimes more, it is not clear how much predictive power it is reasonable

to attribute to such variables under these circumstances, nor whether failure should indict likelihoods in general or just ones covering long periods of time.

The approach in this earlier study was to try to predict purchases of autos reported on Wave 2 as a linear arithmetic function of 13 purchase likehihoods (and various socioeconomic variables) obtained on the first wave. The 13 purchase likelihoods referred to automobiles and to 12 other major durables, which are identified in the stub of Table 1.

The methodological issue arose when two methods of making these predictions were compared -- that based on the regression model and that based on the logit model. The regression model is much simpler but its basic assumptions are contradicted by the use of a 0-1 dependent variable, which clearly does not satisfy the assumptions of normality. Further, it does not restrict the estimates to the allowable range for the dependent variable.

The logit model, on the other hand, is designed for precisely such a situation by expressing the probability of purchase as a simple function of the independent variables. Thus, instead of

$$B_A = \sum_{L=1}^{13} a_i L_i + c$$

where B_A is 0 or 1 as the auto is purchased or not, L_i are the likelihoods, a_i their coefficients, and c a constant, we have:

$$\ln \frac{P_A}{1-P_A} = \sum_{L=1}^{13} a_i L_i + c'$$

or:

$$P_A = \frac{e^{\sum a_i' L_i + c'}}{1 + e^{\sum a_i L_i + c'}}$$

where P_A is the probability of purchasing an auto, the L_i are as before, and the primed terms are the same as before, but marked in this way to indicate that they are the results of a different estimation process.

The results of these computations suggested that at least for the second wave both auto likelihood and some of the other 12 likelihoods made significant contributions to the prediction of auto purchases six months later. Income also was highly significant but none of the other socio-economic variables. Hence, with respect to our substantive question we had some support that configurations of likelihoods helped improve predictions of auto purchases. In addition, comparison of the two methods of fit indicated that the logistic technique did somewhat better in terms of goodness of fit and significance of parameter estimates.

Present Objective

The present paper represents an attempt to determine by purely statistical methods the extent to which the preceding results remain valid when applied to the first nine waves of data for this panel, that is, covering a five-year period. The principal question to be answered on the substantive side is whether the significance of the configuration of purchase likelihoods relative to later auto purchases stood up through time. In other words, do the purchase likelihoods for these other products help to predict auto purchases over and above the information contained in the purchase likelihood for autos alone?

It was also of interest to ascertain whether these likelihoods retained significance after inclusion of the principal socioeconomic variables that seemed to influence auto purchases in these panel data, namely, income level, number of children and whether the wife was working.

From a methodological point of view, we were interested in exploring further the relative merits of regression analysis and logit analysis. While logit analysis is theoretically more defensible, if essentially the same results are obtained by the simpler regression analysis a case could be made for its use despite its theoretical deficiencies.

The plan of this exploratory study was to test four different approaches on these data. First, as a start and as a basis for later comparison, auto purchases on Waves 2 through 9 were expressed as a simple function of auto likelihoods on the preceding wave, and the parameters estimated by both regression and logit analysis.

Second, the configuration of likelihoods was introduced by expressing auto purchases on one wave as a function of all 13 purchase likelihoods on the preceding wave as independent variables. In Waves 6, 8 and 9, separate likelihoods were available for both the husband and the wife. Since there was no clear basis for using one set rather than the other, separate functions were fitted using each set. Again the parameters were estimated both by regression and by logit analysis.

Third, an attempt was made to explore the effect of only past purchases of automobiles and auto purchase likelihoods on actual purchases at a later time. This was done for Wave 9, using as independent variables auto purchases on each of the preceding seven waves and also auto likelihood on each of the preceding eight waves. Once more, the parameters were estimated both by regression analysis and by logit analysis.

Fourth, the three socioeconomic variables that seemed especially important from the previous study were added to the functions tested in the third step. These variables were income, number of children, and whether or not the wife was working.

Substantive Results

Estimates of the parameters of the logit function for Waves 2 through 8 for the prediction of auto purchase in Wave t given the purchase likelihoods on the preceding wave are given in Table 1. In addition to the usual measures of significance of the coefficients, two overall measures of goodness of fit are provided. One measure is a goodness-of-fit statistic associated with the logit method, the Likelihood Ratio Index (LRI). The value of LRI, like the value of R^2, ranges between 0 and 1. However, no clear algebraic relationship exists between these two statistics and the values obtained are not comparable. Moreover, LRI, unlike R^2, has no known distribution properties. Its major use lies in making comparisons among logit functions.

The second overall measure of fit presented in Table 1, the Likelihood Ratio (LR) statistic does have distributional properties, being a chi-squared statistic, but is unbounded in range. However, because of its distributional properties, significance estimates can be attached to this statistic, as is done in the table.

As is evident from Table 1, it appears that when we consider the configuration of likelihoods, as a rule the auto likelihood is the only one to be consistently statistically significant; the others crop up infrequently and without apparent pattern. The overall fit is statistically significant in most instances, but this is almost entirely due to the auto likelihood

LOGIT ON PROBABILITY OF BUYING AUTO ON WAVE t (t=2,...,9) AS A FUNCTION OF 13 LIKELIHOODS FROM WAVE t-1 (N=127)

Variable	Wave of Observed Purchase										
	2	3	4	5	6		7	8		9	
					Husband	Wife		Husband	Wife	Husband	Wife
X_1 - Auto likelihood	.385	.725[c]	.647[b]	.820[c]	.333	.238	.541[c]	.865[c]	.697[c]	.701[c]	.923[d]
X_2 - TVBW likelihood	-.471	.130	-.503	.055	-.122	.197	-.177	-.145	-.028	-.087	-.207
X_3 - TVC likelihood	-.611[a]	-.243	-.266	-.384	.261	.077	.067	-.088	.015	-.133	.366[c]
X_4 - Stereo likelihood	.019	-.030	-.109	-.226	-.132	-.073	-.060	.171	-.025	.082	.556
X_5 - Refrig likelihood	.033	-.653[a]	-.361	-.080	-.212	-.116	.225	.245	.186	-.137	-.035
X_6 - Freezer likelihood	.223	.123	.331	.138	.079	-.126	.258	-1.197	.052	-.237	-.290
X_7 - Stove likelihood	-.755[a]	.609[a]	-.193	-.048	.180	.180	.055	-.991	-.387	-.610	-.305
X_8 - Wash.Mach. likelihood	1.053[a]	.080	-.208	.887[b]	.099	-.081	-.519	-.543	-.282	.943	-.183
X_9 - Dryer likelihood	-1.086[a]	.063	-.130	-.611	-.579	-.441	.286	.290	.445	.338	.392
X_{10} - AC-RM likelihood	.465[a]	.296	-.333	-.607[b]	-.290	.084	-.101	-.056	.094	-.532	.077
X_{11} - AC-Central likelihood	-.504	-.143	-1.016	.179	.078	.465[b]	-.081	-1.887	-.157	.536[a]	.006
X_{12} - Dish.Wash. likelihood	-.834	-.489	-.242	.050	-.209	.218	-.099	-24.989	-.044	-.430	-.120
X_{13} - Was. Dis. likelihood	.288	.432	-.467	-.369	.187	-.166	.051	-13.627	-1.169	.062	-.850
C - Constant	-1.890[d]	-1.505[d]	-2.225[d]	-1.307[d]	-.274	-.251	-.017	-12.021	-1.855[d]	-.458[a]	-.525[b]
LRI	.41	.34	.46	.30	.10	.10	.09	.48	.40	.22	.23
LR Statistic	73.5[d]	60.7[d]	81.7[d]	54.4[d]	18.6	17.6	15.4	85.4[d]	70.8[d]	39.4[d]	40.8[d]
Auto likelihood only:											
LRI	.27	.27	.36	.23	.03	.01	.05	.33	.35	.08	.13
LR Statistic	48.2[d]	48.9[d]	63.9[d]	40.4[d]	5.1[a]	2.5	8.0[b]	59.8[d]	61.9[d]	14.3[d]	23.9[d]

NOTE 1: For t = 6,8,9 there were two sets of likelihoods available on the previous wave. One set was the husband's and the other the wife's. We ran each separately.

NOTE 2: The likelihoods on wave 8 (going with purchase on wave 9) were estimates for the forthcoming 6 months, whereas all others were applicable to the forthcoming 2 years.

[a] Significant at .10 level.

[b] Significant at .05 level.

[c] Significant at .01 level.

[d] Significant at .001 level.

variable. The addition of purchase variables from the preceding wave did not product statistically significant results and are therefore not shown here.

The second approach, relating purchases to all previous purchases and all previous auto likelihoods, indicates a somewhat erratic pattern (Table 2), the most important variable being auto purchase likelihood on the preceding wave. The signs and significant variables suggest a possible three-year buying cycle in cars, but the net contribution of these variables to the explanation of purchases in Wave 9 is relatively small. Moreover, the addition of the three socioeconomic variables, shown in the second column of coefficients in Table 2, contributes somewhat more to the goodness of fit, judging by the increased value of the likelihood ratio statistic. Indeed, comparison of the likelihood ratio statistics for various combinations of variables shows that these three variables (particularly number of children) do make a statistically significant contribution to the goodness of fit, as does the set of past purchase variables.

From this, we can draw two sets of conclusions. First, our answers to whether configurations of likelihoods are important and whether they form patterns through time are both in the negative. We cannot conclude, as before, that the whole configuration is important. In fact, only auto likelihoods seem relevant. Second, the configuration of past purchases seems to have an effect on present ones, the immediately previous purchase by itself does not, but purchases two or three years earlier do seem to influence current purchases.

2. LOGIT COEFFICIENTS ON PROBABILITY OF BUYING AUTO ON WAVE 9 AS A FUNCTION OF PREVIOUS PURCHASES, PREVIOUS LIKELIHOODS, AND OTHER VARIABLES (N=129)

Symbol	Variable	Estimated coefficient 1	Estimated coefficient 2
B_2	Buying - wave 2	.011	.034
B_3	Buying - wave 3	.425	.561
B_4	Buying - wave 4	.503[a]	.639[b]
B_5	Buying - wave 5	-.656[a]	-.942[b]
B_6	Buying - wave 6	-.238	-.308
B_7	Buying - wave 7	.227	.349
B_8	Buying - wave 8	-.412	-.454
LA_1	Auto likelihood - wave 1	-.509	-.895[b]
LA_2	Auto likelihood - wave 2	.694[a]	1.100[a]
LA_3	Auto likelihood - wave 3	-.018	.224
LA_4	Auto likelihood - wave 4	-.594[a]	.665[a]
LA_{5H}	Auto likelihood - wave 5	.313	.203
LA_{5W}	Auto likelihood - wave 5	-.978[b]	-1.167[b]
LA_6	Auto likelihood - wave 6	-.137	-.361
LA_{7H}	Auto likelihood - wave 7	.393	.653
LA_{7W}	Auto likelihood - wave 7	.232	.220
LA_{8H}	Auto likelihood - wave 8	.230	.291
LA_{8W}	Auto likelihood - wave 8	1.218[c]	1.487[d]
IN69	Income 1969		.522
KIDS 73	No. children 1973		-.644[b]
WWORK 73	Wife working 1973		.100
	C - Constant	-.602[b]	-.774
	LRI	.34	.40
	LR Statistic	60.8[d]	70.7[d]

[a]Significant at .10 level.

[b]Significant at .05 level.

[c]Significant at .01 level.

[d]Significant at .001 level.

Logit versus Regression

Regression results corresponding to the logit functions shown in Table 1 yield a very similar pattern of significance, as is shown in Table 3. As before, the auto likelihood variable for the preceding wave is almost invariably statistically significant at the .05 level or beyond, while the other purchase likelihoods are statistically significant only occasionally and not with any particular pattern. Moreover, the net contribution of these other variables to the auto likelihood variable is not statistically significant.

The coefficients estimated by the two procedures (exclusive of constants) have practically all the same signs. They differ only in several cases which are not statistically significant. A few differences do arise with respect to significance. In Tables 1 and 3, the logit and regression results have 16 variables in common that are significant, and each one contributes four variables that are significant and not matched by the other.

In a similar manner, the results from the regression analysis using Wave 9 purchases as a function of previous purchases, previous auto likelihoods and other variables are very similar to the logit results shown in Table 2. The same past-purchase and purchase-likelihood variables are significant in the regression equations (Table 4) as in the logit equation. The principal difference is that with the regression function including the three socioeconomic variables, the income variable is statistically significant at the .10 level while the variable for the number of children is not statistically significant. For both types of methods, however, the addition of the socio-economic variables produces a moderate increase in the goodness of fit.

3.? REGRESSION OF PURCHASE OF AUTO ON WAVE ? (...) ON 13 LIKELIHOODS FROM WAVE 1-2 (N = ...)

ESTIMATES OF COEFFICIENTS

Variable	Wave of Observed Purchase										
	2	3	4	5	6		7	8		9	
					Husband	Wife		Husband	Wife	Husband	Wife
1 – Auto likelihood	$.061^a$	$.107^c$	$.083^b$	$.149^d$	.074	.052	$.125^c$	$.120^c$	$.122^d$	$.131^c$	$.178^d$
2 – TVBW likelihood	−.060	.024	−.043	.012	−.027	.046	−.041	−.0004	−.005	−.012	−.038
3 – TVC likelihood	$-.070^a$	−.043	−.025	$-.066^a$	.058	.016	.016	.015	.010	−.018	$.071^a$
4 – Stereo likelihood	−.013	−.008	−.016	−.040	−.030	−.017	−.012	.021	.002	.007	.071
5 – Refrig likelihood	.003	$-.096^a$	−.027	−.012	−.042	−.024	.051	.036	.030	−.013	−.006
6 – Freezer likelihood	.029	.019	.044	.022	−.020	−.028	.058	$-.070^b$	.0005	−.040	−.056
7 – Stove likelihood	$-.093^a$	.084	−.025	−.007	.035	.037	.009	−.046	−.048	−.068	−.048
8 – Wash.Mach. likelihood	$.126^a$	.005	−.026	$.139^b$	.011	−.015	−.107	−.027	.0006	.070	−.037
9 – Dryer likelihood	$-.140^b$	.024	−.010	−.090	−.116	−.096	.056	.005	.022	.042	.070
10 – AC-RM likelihood	.071	.047	−.035	$-.096^b$	−.061	.018	−.021	−.035	.010	−.051	.006
11 – AC-Central likelihood	−.028	−.020	−.057	.031	.019	.104	−.016	−.064	−.021	.059	.034
12 – Dish.Wash. likelihood	−.053	−.054	−.014	.005	−.046	$.043^b$	−.021	−.020	−.011	−.080	−.009
13 – Was. Dis. Likelihood	.026	.062	−.020	−.062	.042	−.034	.011	−.028	−.043	.040	−.066
C – Constant	$.209^d$	$.232^d$	$.178^d$	.255	$.442^d$	$.442^d$	$.496^d$	$.193^d$	$-.194^d$	$.396^d$	$.396^c$
R^2	$.17^c$	$.15^a$	.15	$.17^b$	.12	.16	.11	$.18^b$	.14	$.20^c$	$.23^c$
R^2_{adj}	.08	.06	.06	.08	.02	.07	.01	.09	.04	.11	.14
Auto likelihood only:											
R^2	.01	$.08^c$	$.05^b$	$.06^c$	$.02^a$	.006	$.06^c$	$.07^c$	$.09^d$	$.07^c$	$.14^d$
R^2_{adj}	--	.07	.04	.05	.01	--	.05	.06	.08	.06	.13

[a] Significant at .10 level.
[b] Significant at .05 level.
[c] Significant at .01 level.
[d] Significant at .001 level.

4. REGRESSION OF AUTO PURCHASE ON WAVE 9 ON PREVIOUS AUTO PURCHASES, PREVIOUS AUTO PURCHASES, PREVIOUS AUTO LIKELIHOODS, AND OTHER VARIABLES

Symbol	Variable	Estimated coefficient 1	Estimated coefficient 2
B_2	Buying wave 2	.003	-.0003
B_3	Buying wave 3	.052	.060
B_4	Buying wave 4	.085[a]	.092[b]
B_5	Buying wave 5	-.085[a]	-.086[a]
B_6	Buying wave 6	-.021	-.024
B_7	Buying wave 7	.035	.043
B_8	Buying wave 8	-.065	-.067
LA_1	Auto likelihood wave 1	-.077[a]	-.096[b]
LA_2	Auto likelihood wave 2	.099[a]	.117[b]
LA_3	Auto likelihood wave 3	-.016	-.002
LA_4	Auto likelihood wave 4	-.076	-.077
LA_{5H}	Auto likelihood wave 5	.044	.034
LA_{5W}	Auto likelihood wave 5	-.150[b]	-.151[b]
LA_6	Auto likelihood wave 6	-.018	-.046
LA_{7H}	Auto likelihood wave 7	.039	.062
LA_{7W}	Auto likelihood wave 7	.043	.034
LA_{8H}	Auto likelihood wave 8	.045	.054
LA_{8W}	Auto likelihood wave 8	.168[d]	.181[d]
IN69	Income 1969		.074[a]
KIDS73	No. Children 1973		-.064
WWork73	Wife Working 1973		.016
C	Constant	.395[d]	.395[c]
	R^2	.35[d]	.39[d]
	R^2_{adj}	.24	.27

[a]Significant at .10 level.

[b]Significant at .05 level.

[c]Significant at .001 level.

We do find an almost universal tendency for the LRI to be greater than R^2, both for Tables 1 and 3, and for Tables 2 and 4. It is not clear what, if anything, this shows. More important is that the LR statistic both shows slightly higher levels of significance and is significant more often than R^2. This suggests a better fit from the logit method, although how much is not clear.

To obtain an idea of the relative prediction abilities of these two approaches, each of the functions was used to derive an error classification matrix, that is, to actually "predict" the auto purchase of each household so that the accuracy of classification of each function could be assessed. Since this test is applied in the present case to the same observations from which the parameters were estimated, the possibility of search bias is quite high, and it is a pity that not enough observations were available to enable the sample to be split into separate analysis and validation samples. Nevertheless, these results should provide some indication of the extent if any to which these two methods differ.

The results of these computations are shown in Table 5. Rather surprisingly, they show that, despite the apparent similarity of the two sets of results in terms of significant coefficients and goodness-of-fit, the logit model has an accuracy of classification either approximately equal to, or far superior than, the regression model. This is especially true for the Wave 9 functions that combine previous auto likelihoods with previous auto purchases.

Equally interesting is the fact that, contrary to the results obtained with the goodness-of-fit measure, higher accuracy of classification is obtained for some of the earlier wave functions using only the 13 purchase likelihoods or only sequences of previous auto purchase likelihoods than the more complex Wave 9 functions. This is true both of the logit model and of

5. PERCENT OF AUTO PURCHASE REPORTS CLASSIFIED CORRECTLY BY ALTERNATIVE LOGIT AND LINEAR REGRESSION MODELS, BY WAVE

Model	Wave*	Logit	Regression
All 13 likelihoods	2	82.9%	78.2%
	3	79.8	76.7
	4	82.2	82.9
	5	76.0	69.0
	6H	63.6	41.7
	6W	65.9	45.0
	7	65.9	34.9
	8H	85.3	79.0
	8W	82.2	77.5
	9H	73.6	55.8
	9W	72.1	55.0
Previous auto likelihoods	2	78.3%	51.9%
	3	79.1	79.8
	4	82.2	82.9
	5	74.4	75.2
	6H	53.5	41.1
	6W	55.8	56.6
	7	59.7	30.2
	8H	80.6	81.4
	8W	79.8	81.4
	9H	68.2	58.1
	9W	70.5	56.6
Previous auto likelihoods and auto purchases	9	74.4%	52.7%
Previous auto likelihoods and auto purchases and socioeconomic variables	9	78.3%	51.9%

*The H's and W's refer to which of the two sets of likelihoods (i.e. husband's or wife's) available on the wave were used as independent variables.

the regression model. However, for Wave 9, which had the most combinations tested, the accuracy of classification of the logit model (but not of the regression model) is highest when auto likelihood and auto purchase are combined with the socio-economic variables. Especially interesting is the fact that the actual percent purchasing autos on Wave 9 was 40%, so that the logit results but not the regression results are far better than what would be obtained by a naive model estimate.

These results should be treated with caution since they are obtained from the same observations used to estimate the parameters of the models. Nevertheless, they support the finding obtained many times in the past of the unreliability of goodness-of-fit as a measure of predictive accuracy. Moreover, they leave the strong supposition that the logit model is no worse than the regression model and may be considerably better. Indeed, if the logit model is as accurate in classifying other types of observations as it is for these data, it would seem to provide a very useful forecasting tool both for micro and macro purposes.